# 18 Statistics Primer

Learning Objectives

* Understand the purpose of statistics
* Explain descriptive statistics and its methods
* Know the various measures of central tendency and spread in data
* Learn about histograms
* Learn about normal distribution and its usefulness
* Show how to draw inferences from a random sample

### INTRODUCTION

Statistics is a mathematically elegant and efficient way of understanding the characteristics of the whole population by projecting from the understanding of a sampled part. Ideally, the statistical analysis would give a completely ac- curate and prudent description of the whole population with the least amount of effort. This is achieved by accessing a truly random representative sample of the population, gathering data on sample observations, analyzing the data to discover patterns, and projecting the results with a certain high minimum level of confidence. Statistical analyses ultimately lead to a deeper understanding of the Totality, the unified field of all the laws of nature.

There are many applications of statistics. It is most useful in understanding human populations. Politicians conduct polling surveys to understand the needs of voters. Marketers use statistics to understand the needs and wants of consumers. Statistics are also useful in assessing quality in production processes by measuring the quality of a sample of products.

### DESCRIPTIVE STATISTICS

These are the tools and techniques to describe a collection of data. Data is often described by its central tendency and its spread. The primary central tendency is the *mean*, or the average, of the values. There are other central tendencies such as the *median* and *mode*. The spread within the data is called *variance* and is often described by *‘standard deviation’*.

### Example Dataset

For example, use the data of people as shown in Dataset 18.1 given below.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset 18.1 |  | | |
| Gender | Age | Height | Weight |
| M | 24 | 71 | 165 |
| M | 29 | 68 | 165 |
| M | 34 | 72 | 180 |
| F | 21 | 67 | 113 |
| M | 32 | 72 | 178 |
| F | 25 | 62 | 101 |
| M | 26 | 70 | 150 |
| M | 34 | 69 | 172 |
| M | 31 | 72 | 185 |
| F | 49 | 63 | 149 |
| M | 30 | 69 | 132 |
| M | 34 | 69 | 140 |
| F | 28 | 61 | 115 |
| M | 30 | 71 | 140 |

A person might be described by his/her gender, age, height, and weight, for example. Those are the attributes of a person. A collection of persons can be described by their total count and their average age, height, and weight. The averages could be computed separately for different values of gender.

Averages need to be complemented by the spread or variance in the values of the attributes. For example, two groups may have the same average age. However, in one group, all the members have the same age. While in the other group, the age ranges from low values to high values. In the first group, the variability, or spread, is zero. In the other group, there is a non-zero spread.

### Computing Mean, Median, Mode

The count of instances in this dataset is 14. The average age is 30.5 years. (It can be computed by using the function ‘Average’ in MS Excel.)The average height is

68.3 inches and the average weight is 149.9 lbs.

To appreciate the concept of spread, it helps to sort the data by one of the at- tributes. This data is sorted by ‘Age’ and shown below in Dataset 18.2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | | |  |
| Dataset 18.2 |  |  |  |  |
| Gender | Age | Height | Weight |  |
| F | 21 | 67 | 113 |  |
| M | 24 | 71 | 165 |  |
| F | 25 | 62 | 101 |  |
| M | 26 | 70 | 150 |  |
| F | 28 | 61 | 115 |  |
| M | 29 | 68 | 165 |  |
| M | 30 | 69 | 132 |  |
| M | 30 | 71 | 140 |  |
| M | 31 | 72 | 185 |  |
| M | 32 | 72 | 178 |  |
| M | 34 | 72 | 180 |  |
| M | 34 | 69 | 172 |  |
| M | 34 | 69 | 140 |  |
| F | 49 | 63 | 149 |  |

The age of the members in this group ranges from 21 years to 49 years. The average may not tell the full story about the group. Thus, another way to describe the central tendency of the age of this group is to describe the age of the person in the middle of this group. The age of the middle member in this group of 14 people is 30 years. That is called the median value of the age of the group. (It can be computed by using the function ‘Median’ in MS Excel.) Median is particularly useful when the distribution of values is highly skewed. Thus, the median height of the group is 69 inches and the median weight is 149.5 lbs.

Another way of describing the central tendency is of particular relevance when the values are non-numeric, such as their gender, native language, etc. It is not possible to average out the non-numeric data. Thus the concept of ‘mode’ becomes important. The mode is the value that shows the highest frequency of occurrence in the data. In this data, the age value of 34 occurs three times, the age value of 30 occurs two times, and all the other values occur just one time each. Thus the mode value of the age of this group is 34 years. (It can be computed by using the function ‘Mode’ in MS Excel.) Similarly, the modal height of the group is 69 inches and 72 inches as both occur 3 times each. In such cases, it is called bi-modal data. The modal value of weight is 140 lbs.

The concept of mode can be useful in describing the dominant gender of this group. The M value occurs 10 times and the F value occurs 4 times. Thus one can say that the modal value of the gender for the group is M.

### Computing the Range and Variance

The spread of the data can be defined in two primary ways. One is the range, and the other is variance.

The range is simply defined by the minimum and maximum values of the variable. Thus, the range of values of age in this dataset is from 21 to 49 years. Similarly, the range of height is from 61 to 72 inches, and the range of weight is from 101 to 185 lbs. The greater the range, the more there is variation in the data; the lower the range, the greater is the homogeneity in the data. The limitation of the range measurement is that it depends only upon the extreme values on either side and ignores the distribution of the rest of the data in the middle.

The variance is a more inclusive measure of the spread in the data. Variance is defined as the sum of squares of the distance from the mean. Thus, to compute the variance for the age variable for this group, first, calculate for each instance, the difference between the age and the average value for the group (which is 30.5). Make a square of those difference values by multiplying each value with itself. Summing up those squared values leads to the total variance. Dividing the total variance by the number of instances leads to the average variance of the data. The standard deviation is defined as the square root of the average variance. Mathematically speaking, given below is the formula for standard deviation

*s* =

S(*x* - *x* )2

*n* - 1

where *S* is the standard deviation. S means the sum of all the values. is the

value of the variable for an instance in the dataset. is the mean for that variable in the dataset. *n* is the number of instances in the dataset.

Thus computing the average variance for age in this group will lead to a value of 6.64 years. (It can be computed using the function ‘Stdev’ in MS Excel). Similarly, the standard deviation can be computed for all other numeric variables. The standard deviation for height is 3.75 years, and the standard deviation for weight is 26.8 lbs.

Thus, any dataset can be defined quite meaningfully in terms of its mean and standard deviation.

### Histograms

Histograms are a particularly useful way of visually describing the distribution of data. Given below are histograms for each of the variables.

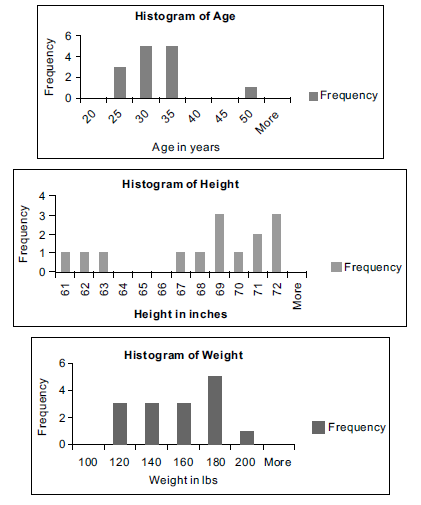


FIGURE 18.1 Histograms of Different Variables

Histograms are usually developed for numeric variables, though in principle they can also be developed for non-numeric variables such as gender.

These histograms show a wide range of data values with major gaps in the middle. Therefore, the range of each variable should be specified in addition to the mean value to give a more accurate picture of the data. And more importantly, the gaps suggest that there are probably different subgroups within the data.

In this case, likely, the values vary significantly based on the value of gender. The Analysis of Variance (ANOVA) technique helps determine the impact of such variables.

### NORMAL DISTRIBUTION AND BELL CURVE

The data is normally distributed if the data values are evenly distributed like a symmetric bell-shaped curve.

–3

–2

–1

1

2

3

FIGURE 18.2 A Symmetric Bell-shaped Curve

The central tendency is the middle point of the distribution in data. The perfect symmetry ensures that this central point is the mean, as well as the median, and the mode for the data distribution. The frequency of the values decreases on either side of the central value. The numbers on the *x*-axis represent the number of standard deviations away from the central tendency, or the mean. Approximately 95 percent of the data values, the area under the curve, fall within 2 standard deviations (SD) on each side.

If the dataset were normally distributed, 95 percent of the age values in the dataset will fall within mean ±2\* SD. Assuming the age is normally distributed, 95 percent of the data values would fall within that 2 SD range, i.e., (30.5 – 2 ¥ 6.64) to (30.5 + 2 ¥ 6.64), or from about 17 to 44 years. This adequately reflects this dataset.

A normal distribution assumes a natural clustering of values around the central value. It is considered a naturally occurring pattern in most situations. In most datasets, data is assumed to be normally distributed unless otherwise specified. As the dataset gets larger, it would more and more begin to resemble the bell curve.

The normal distribution can be mathematically modeled in terms of two parameters – mean and standard deviation. Parametric curves have many properties and benefits, for example, normal curves can be instantly compared. Also, the area under the normal curve can be used using calculus functions and will follow another symmetric curve called the *S*-curve. That can be used in further computations. A normal distribution of data helps infer the properties of the population from a randomized sample with a certain high degree of confidence.

### INFERENTIAL STATISTICS

The primary purpose of statistics is to infer the properties of the population without having to reach out to and measure the entire population. If a suitably randomized sample is selected, such that it can reasonably be assumed to be representative of the overall population, then the properties of the population can be inferred from it with a certain high degree of confidence. The larger the sample size, the greater would be the confidence.

### Random Sampling

A random sample is a mathematically rigorous technique of selecting a sample that is representative of the population. It should be contrasted with other forms of sampling, such as convenience sampling, which finds the easiest way to gather data without caring much about its representativeness of the entire population. The random sampling process uses a random number generating mechanism to select a random from a certain number of instances from the total population, such that every instance has an equal chance of being selected. An example of random selection would be how lottery winners are selected from the entire pool of lottery tickets sold. The size of the random sample is chosen by the statistician based on the accuracy needs and the time and resources available.

The random sample can be analyzed using any of the descriptive techniques described above. And even analytical techniques such as Analysis of Variance.

### Confidence Interval

Unless the entire population has been measured, a completely true value of the variable for the population cannot be provided. The random sampling method can provide an estimate of the true mean of the population, with a certain degree of confidence. The true population means will fall within a range centered on the sample. This is called the confidence interval. The larger the sample size, the narrower will be the confidence interval. The greater the sample size, the greater will be the confidence in the projection from the sample mean.

Suppose that the dataset used above was representative of the people in a certain community. How confidently can one claim that the average age of the population of the community is 30.5 years (the average age of the sample)?

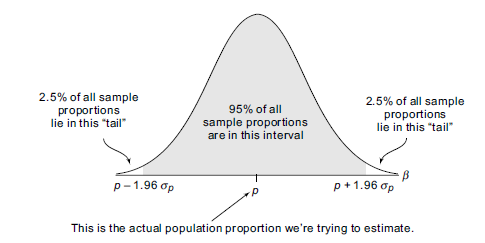


FIGURE 18.3 Determining the Confidence Interval

The range, around the sample, mean, within which the true population mean (*P*) is likely to fall, is called a confidence interval. The higher the confidence level that is desired, the wider will be the confidence interval.

Based on the normal distribution of data, the confidence interval of 95 percent translates into 2 standard deviations worth of distance on either side of the mean. It means that if such randomized sampling and analysis were done often, then 95 percent of the times the confidence interval of mean ±2 SD will bracket (i.e., include) the true population means. The confidence level is usually specified by its inverse, the chance or error, i.e., 1 – a, where a is called the significance level, or simply the *p*-value. For a 95 percent confidence level, the *p*-value should be less than or equal to 0.05.

### PREDICTIVE STATISTICS

Statistical techniques such as regression analysis can be used for predicting the values of variables of interest. Correlations among variables show which variables might influence the values of others. There are many variations of regression, such as linear, non-linear, and logit regression. Time series analysis is a special

case of regression analysis where the key independent variable is time. These are described in full detail in chapter 7.

Conclusion

Statistical tools are time-tested ways of understanding large populations using small randomized samples. Data is described using central tendencies such as mean, median, and mode; and its distribution using range and variance. Predictive statistical methods like regression and others are prominent tools in data analytics. These are fully covered in the chapter on Regression.

## Questions

1. What is statistics? What are its different types?
2. What is a histogram? How does it help?
3. What is the difference between central tendency and spread?
4. What is the difference between mean, mode, and median?
5. What is the standard deviation? How is it computed? How does it help?
6. What is random sampling? How does it help?
7. What is the difference between confidence level and confidence interval?

## True/False

1. Statistics is an elegant and efficient way of counting averages.
2. The mean value is almost always lower than the median value.
3. There can be multiple modal values in a dataset.
4. Standard deviation is a measure of the count of instances in the dataset.
5. The range of a dataset is defined by its mean and standard deviation.
6. A convenience sample helps conduct statistical research faster and more effectively.
7. The larger the sample size, the narrower the confidence interval.
8. A histogram is a graphical device to show the frequency of data distribution.
9. Almost all data in nature is normally distributed.
10. MS Excel is not suitable for doing statistical analysis.